MUG: A General Meeting Understanding And Generation Benchmark

Qinglin Zhang[1] , Chong Deng[1] , Jiaqing Liu[1] , Hai Yu[1] , Qian Chen[1] , Wen Wang[1] , Zhijie Yan[1] , Jinglin Liu[2] , Yi Ren[2] , Zhou Zhao[2]

1 Speech Lab of Damo Academy, Alibaba Group, China
2 Zhejiang University, China

ICASSP 2023

Introduction

- **Background**
  - NLP applications on meeting transcripts significantly enhance users' efficiency in grasping important information

- **Challenges**
  - Lack of large-scale public meeting datasets with spoken language processing (SLP) annotations

  - Meeting transcripts pose great challenges to SLP compared to written and formal text
    - Exhibit a wide variety of spoken language phenomena
    - Typically lengthy documents (several thousand words or more)
    - ASR errors further drastically degrade SLP performance

Introduction

- **Prior Meeting Datasets Supporting SLP Development**
  - The ICSI meeting corpus
  - The AMI meeting corpus
  - The ELITR Minuting Corpus
- **Our Goal**
  - Establish a General Meeting Understanding and Generation (MUG) Benchmark
  - Construct and release a large-scale meeting dataset – the AliMeeting4MUG Corpus with representative and diverse SLP annotations on manual transcripts
  - Prompt SLP research on meetings

Dataset Collection and Annotations

- **Our AliMeeting4MUG Corpus (AMC)**
  - To the best of our knowledge, AMC is so far the **largest meeting corpus** and **facilitates most SLP tasks**
  - **654** meetings, 15-minute to 30-minute discussions by 2-4 participants, diverse topics, biased towards work meetings
  - Manual transcripts with manually inserted punctuation and speaker labels
  - Manual annotations for 5 SLP tasks

Datasets	#Sessions	#Avg. Turns	#Avg. Speakers	Avg. Session Len.	Supported Tasks	Language
AMI	137	535.6	4.0	5,570.4	Action, SUM, TS	English
ICSI	59	819.0	6.3	8,567.7	Action, SUM, TS	English
ELITR (English)	120	727	5.9	7,066	SUM	English
ELITR (Czech)	59	1,205	7.6	8,534	SUM	Czech
QMSum	232	556.8	9.2	12,026.3	QA, SUM, TS	English
AMC (ours)	**654**	376.3	2.5	10,772.5	**Action, KPE, SUM, Title, TS**	Mandarin

Access the AMC corpus:

https://www.modelscope.cn/datasets/modelscope/Alimeeting4MUG/summary

AMC: Data Collection and Annotations

達摩院 ALIBABA DAMO ACADEMY

AMC Example

.....
参会人1：客服销售得来一个手机吧一个人，或者是给报销一下话费。
参会人0：嗯。
参会人2：那要每个人来个手机话，这样花销也太大，咱公司，财务那边肯定就不批。
参会人0：是啊这。
参会人3：你别说销售，销售他是他是有公用电话的，每工位都有公用电话，那给他配手机干嘛使啊？
参会人0：对。那销售他们用那个固话打嘛，他们销售销售就用，他们那个销售固话打嘛。
参会人2：对对。
参会人3：对呀。 disfluency
参会人1：但是那个销售私自联系呃那个私下联系客户的话，不可能会用那个公用电话，他下了班之后依旧在工作呀。
参会人3：财务也有固话。
参会人0：他。所以咱们咱们可以这样嘛，公司每个月就补贴一点话费。
参会人3：呃，下了班下了班以后，他们可以在公司加班儿呀。对吧，他们可以，客户约到几点几点，可以用公司电话打呀。 disfluency
参会人0：是不是可以这样啊，咱们每个月就是不给给员工报，对对对。就是，你说一个月打到五百是吧，一个月打到五百公司肯定不可能都报。一个月就补贴一点就行了。
参会人2：还是给补助点儿话费比较合适，对给补助点儿话费。
参会人1：补助点话费吧给。
参会人3：行行行。 grammar error
参会人2：对。
参会人1：对如果说十几二十块的话，就不要报销了，五百多那得报销一下吧。
参会人2：对。
参会人0：就，就每个月每个月跟个，每个月咱们就就就。
.....

AMC: Dataset Collection and Annotations

達摩院
ALIBABA DAMO ACADEMY

QMSum Example

.....
Grad F : Mm - hmm .
PhD H : So you have spare headsets ?
Postdoc A : Sorry , what ?
PhD H : You have spare headsets ?
Grad F : They 're just earphones . They 're not headsets . They 're not microphones .
PhD E : Right .
PhD H : No , no . I mean , just earphones ? Um , because I , uh , I could use one on my workstation , just to t because sometimes I have to listen to audio files and I don't have to b go borrow it from someone and
Postdoc A : We have actua actually I have W Well , the thing is , that if we have four people come to work for a day , I was I was hanging on to the others for , eh for spares ,
PhD H : Oh , OK
Postdoc A : but I can tell you what I recommend .
Professor B : No , but you 'd If you Yeah , w we should get it .
PhD H : Sure . No problem .
.....

disfluency

disfluency

disfluency

AMC: Dataset Collection and Annotations

达摩院
ALIBABA DAMO ACADEMY

- 5 Spoken Language Processing Annotations
  - Topic Segmentation (TS)
    - Label only the last sentence of paragraphs
  - Extractive Summarization (ES)
    - Label key sentences for each topic and for each session respectively
  - Topic Title Generation (TTG)
    - Create an informative and concise title for each topic by summarizing its central idea
  - Keyphrase Extraction (KPE)
    - Label top-K keyphrases for a session
  - Action Item Detection (AID)
    - Label sentences containing information about actionable tasks

AMC: Exploring Multi-Annotator Annotations

- **Annotation Process**
  - TS
    - One annotator annotates and another expert reviews/corrects labels
  - ES, TTG, KPE, and AID
    - Three annotators
    - AID: Another expert reviews and decides the final label for training&evaluation
- **Inter-Annotator Agreement (IAA)**
  - ES and TTG: ROUGE-1,2,L F-score
  - KPE: Exact F1
  - AID: Kappa coefficient
- **Training and Evaluation Labels**
  - ES: Union of labeled sentences for training, report avg. and best ROUGE scores based on three annotations for evaluation
  - TTG: Copy and pool for training, report avg. and best ROUGE scores based on three annotations for evaluation
  - KPE: Union of labels for training and evaluation

AMC: Exploring Multi-Annotator Annotations

Data Statistics and IAA for All SLP tasks on AMC

	TS		Topic-level ES	Session-level ES	TTG	KPE	AID
IAA	N/A		49.53/30.50/41.13	55.65/28.40/34.97	30.79/16.63/28.17	55.62	0.50
	#Topics	Len.	Count/Topic	Count/Session	Len.	Count/Session	Count/Session
mean	9.81	996.1	2.41	10.81	11.26	17.37	3.22
std	2.22	353.9	0.66	2.93	1.85	3.53	3.86
25%	8	714	2	9	10	15	0
50%	9.5	950	3	10	11	17	2
75%	11	1230	3	12	13	20	5

Observation: **The moderate IAA values indicate great challenges of SLP annotations on meetings**, which demand more studies.

Task Setting and Evaluation Metrics

達摩院
ALIBABA DAMO ACADEMY

Tasks	Task Definition	Evaluation Metrics
Topic Segmentation (TS)	Segment transcripts of a session into a sequence of non-overlapping topically coherent segments	Positive F1, Pk, and Win-Diff (WD)
Extractive Summarization (ES)	Extract key sentences for each reference topic segment and the entire session, without modifying original sentences	Average and best ROUGE-1,2,L
Topic Title Generation (TTG)	Generate an informative and concise title for each reference topic segment	Average and best ROUGE-1,2,L
Keyphrase Extraction (KPE)	Extract top-K keyphrases from a session that can reflect its main content	Exact F1 and Partial F1
Action Item Detection (AID)	Detect sentences containing information about actionable tasks as positive samples	Positive F1

Baseline Systems: Model Selection

- **TS. ES, AID**: Longformer-base as backbone
  - Better at handling long-form document with linear complexity
  - Window-based self-attention to capture local context
  - Task-specific global attention to encode inductive bias about the task
- **TTG**: BART-base as backbone
  - Denoising autoencoder for seq2seq modeling
  - Achieve SOTA results on a number of text generation tasks
- **KPE**: YAKE
  - Perform relatively stable on documents with varying lengths, especially on long documents

Baseline Systems:

https://github.com/alibaba-damo-academy/SpokenNLP/tree/main/alimeeting4mug

Baseline Systems: Results on MUG Test sets

Track 1 Topic Segmentation (TS)			
Model	positive F_1	$1 - p_k$	1-WD
Longformer	$22.7_{\pm 0.98}$	$0.583_{\pm 0.008}$	$0.56_{\pm 0.008}$

Track 2 Extractive Summarization (ES) (AVG)			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$53.83_{\pm 0.39}/61.64_{\pm 0.68}$	$32.33_{\pm 0.60}/42.73_{\pm 0.84}$	$42.94_{\pm 0.61}/53.87_{\pm 0.68}$
Topic-level ES			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$51.16_{\pm 0.68}/63.0_{\pm 1.03}$	$34.4_{\pm 0.78}/49.61_{\pm 1.19}$	$45.03_{\pm 1.02}/59.61_{\pm 1.2}$
Session-level ES			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$56.5_{\pm 0.94}/60.28_{\pm 1.2}$	$30.26_{\pm 0.77}/35.85_{\pm 1.07}$	$40.84_{\pm 0.53}/48.13_{\pm 0.43}$

Track 3 Topic Title Generation (TTG)			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
BART	$32.16_{\pm 0.21}/45.11_{\pm 0.22}$	$17.87_{\pm 0.22}/28.26_{\pm 0.32}$	$30.1_{\pm 0.26}/43.16_{\pm 0.22}$

Track 4 Keyphrase Extraction (KPE)			
Model	Exact/Partial F_1@10	Exact/Partial F_1@15	Exact/Partial F_1@20
YAKE	15.2/24.9	17.5/27.8	19.1/29.5

Track 5 Action Item Detection (AID)			
Model	positive P	positive R	positive F_1
Longformer	$60.18_{\pm 5.06}$	$66.89_{\pm 3.29}$	$63.14_{\pm 1.41}$

Benchmark URL: https://www.modelscope.cn/leaderboard/27/summary

Baseline Systems: Compare to Other Datasets

Topic Segmentation

Datasets	Positive F1 ↑	1-Pk ↑	1-WD ↑
MUG (Meeting Human transcripts)	**21.00**	**0.571**	**0.545**
QMSUM (Meeting ASR transcripts)	21.92	0.675	0.657
wiki-727 (Written Text)	75.45	0.853	0.842

Abstractive Summarization

Datasets	SOTA (ROUGE-L) ↑
MUG (Meeting Human transcripts)	**30.1**
CLES (Written Text)	41.055
LCSTS (Written Text)	48.46

Observations

- With **same baseline systems**, TS performance on AMC manual transcripts is worse than that on QMSUM ASR transcripts
- Our baseline TTG performance on AMC manual transcripts is worse than abstractive summarization SOTA on ASR transcripts of AMI,ICSI and QMSUM meeting corpora and worse than SOTA on written text

- **SLP tasks on AMC could be more challenging compared to on other meeting corpora**
- **SLP tasks on AMC are much more challenging than on written text**

URLs of AMC Data and Our Code

達摩院
ALIBABA DAMO ACADEMY

Download the AMC data

Baseline Systems

Conclusion and Future Work

- **Conclusion**
  - Establish a general and comprehensive Meeting Understanding and Generation benchmark (MUG) to prompt spoken language processing (SLP) research on meetings
  - Construct the AliMeeting4MUG Corpus (AMC) for MUG
    - To the best of our knowledge, AMC is so far the largest meeting corpus and facilitates most SLP tasks
    - Define tasks, conduct SLP annotations, build and evaluate baseline systems

- **Future Work**
  - Release ASR 1-best to prompt research on SLP robustness to ASR errors
  - Add tasks such as Question Answering and Abstractive Summarization variants
  - Cover more languages such as English
  - Facilitate multi-modality MUG research (such as audio, image, video)

Thanks

Q&A